Gradient Prior Dilated Convolution Network for Remote Sensing Image Super-Resolution

Ziyu Liu, Ruyi Feng, *Member, IEEE*, Lizhe Wang, *Fellow, IEEE*, and Tieyong Zeng

***Abstract*—Super-resolution (SR) aims to recover a high-resolution image from a single or multiple low-resolution images, compensating for the limitations of satellite sensor imaging. Deep convolutional neural networks have made great achievement in remote sensing image SR. In this article, we propose a novel gradient prior dilated convolutional network (GPDCN) for remote sensing image SR, which obtains contextual spatial connections and alleviates structural distortions. The GPDCN comprises a multiscale feature extraction network and a feature reconstruction network. The former employs a double-path dilated residual block with dilation convolution to increase a receptive field, a global self-attention module to detect long-range reliance among image patches, and a gradient propagation network to extract high-level gradient information. The latter uses the mixed high-order attention module to reconstruct the feature by collecting the high-order characteristics of multiple frequency bands. Experiments with the Massachusetts_Roads and 3K VEHICLE_SR datasets demonstrate that the GPDCN outperforms recent techniques concerning both quantitative and qualitative measures.**

***Index Terms*—Attention, dilated convolution (DC), gradient prior, remote sensing super-resolution.**

I. Introduction

IMAGE super-resolution (SR) is a popular topic in remote sensing as it is intended to regenerate high-resolution (HR) images from corresponding low-resolution (LR) equivalents. Obtaining HR remote sensing images has become increasingly meaningful in a variety of real-world applications, including resource management, environmental monitoring [1], construction planning [2], [3], and military investigation [4]. Indeed, HR remote sensing images can contain copious amounts of vital information critical to such applications. However, owing to the limitation of imaging technology, the spatial resolution of a remote sensing image usually fails to meet the accuracy requirements [5], [6]. In addition, a variety of factors can decrease the quality of remote sensing imagery, such as transmission noise and motion blur [7]. To reduce the impact from the imaging process, the most direct method is to equip more precise remote sensors; however, this increases the hardware cost. As a result, there is demand for a practical and effective strategy that tackles the limitations of remote sensors across different practical remote sensing applications.

Unlike multiframe images that achieve a better resolution output by establishing a relationship between a targeted HR image and numerous LR images of the same scene under various circumstances. Single-image super-resolution (SISR) must rely entirely on only one input image without any additional available information; this typically results in an ill-posed issue that contains multiple image reconstruction solutions due to a loss of information. Despite these challenges, numerous SISR approaches have been proposed to date. They can be divided into three broad categories [8]: interpolation, reconstruction, and learning-based algorithms.

Interpolation-based algorithms are the most fundamental approaches to image reconstruction. Classical interpolation approaches, such as nearest-neighbor and bicubic interpolation [9], are widely used today. Nearest-neighbor interpolation chooses the closest pixel value for each location being interpolated, and while the method has a quick execution time, it struggles to provide high-quality outputs. Bicubic interpolation conducts cubic interpolation on two axes, and while it produces smoother results with fewer artifacts than other interpolation-based methods, it is slow. Thus, the interpolation approaches have yielded good results by directly using prior knowledge of natural images, but real-world satellite images with intricate details cause difficulties during the reconstruction process. Simple interpolation-based methods can result in overly smoothed edges when an image size increases.

The vast majority of satellite SR approaches use reconstruction-based methods to rebuild matching HR images by extracting the valuable information in the LR image, and combining some prior knowledge, the reconstruction process is constrained. However, these methods primarily depend on the prior knowledge of the HR images, such as gradient-profile [10], edge [11], and smoothness priors [12]. Therefore, the reconstruction-based methods are usually constrained to hand-crafted features that need manual parameter adjustments. Therefore, it is difficult to use them to handle complicated and changing scenes.

Learning-based algorithms have been proposed as a way to avoid the above problems by establishing end-to-end training

Manuscript received 22 November 2022; revised 1 February 2023; accepted 25 February 2023. Date of publication 6 March 2023; date of current version 28 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grants U21A2013 and 41925007, in part by the Hong Kong Scholars Program under Grant XJ2020025, in part by the Hubei Key Laboratory of Regional Development and Environmental Response (Hubei University) under Grant 2020(B)003, and in part by the National Key R&D Program of China under Grant 2021YFE0203700. *(Corresponding author: Ruyi Feng.)*

Ziyu Liu, Ruyi Feng, and Lizhe Wang are with the Hubei Key Laboratory of Intelligent Geo Information Processing and the School of Computer Science, China University of Geosciences, Wuhan 430074, China (e-mail: ziyuliu@cug.edu.cn; fengry@cug.edu.cn; lizhe.wang@gmail.com).

Tieyong Zeng is with the Department of Mathematics, Chinese University of Hong Kong, Hong Kong (e-mail: zeng@math.cuhk.edu.hk).

Digital Object Identifier 10.1109/JSTARS.2023.3252585

between LR and HR image pairs [13], [14], [15]. Deep learning has some remarkable progress in image SR, for which deep convolution neural networks (CNNs) can extract powerful feature representation capabilities. Dong et al. [16] have proposed a deep learning method for directly learning an end-to-end mapping between LR and HR images. Furthermore, Kim et al. [17] have presented SR methods that uses a deeply recursive convolutional network to improve imaging performance without using additional parameters. Nevertheless, using CNN-based SR models for remote sensing images results in several challenges. First, most CNN-based methods boost performance by setting a very deep model; however, such deep models usually cause high computation and memory costs, limiting their practical applications. Second, the use of dilated convolution (DC) is not used effectively in enlarging perceptive fields and multiscale features. Third, most CNN-based SR models insufficiently and inefficiently use the prior knowledge of images.

Thus, this article proposes using a gradient prior dilated convolutional network (GPDCN) for remote sensing image SR to solve the above issues. The model comprises an end-to-end pixelwise network that combines low-level detail information, high-level semantic information, gradient information, and global contextual information. Furthermore, the framework consists of two phases: feature extraction and feature restoration. The feature extraction section is a deep convolutional network architecture that captures more powerful edge characteristics than other methods using gradient prior information. A double-path dilated residual block (DPDRB) extracts multiscale feature maps from LR and gradient images during the process. In addition, a global self-attention (GSA) module appended to each of the first five DPDRBs considers global context relevance. Then, the frequency-based combination of feature maps with varying frequencies is used to reconstruct the final high-frequency details via the mixed high-order attention module (MHOA) in the restoration section, acted as the feature restoration section. In summary, this article makes the following contributions.

1) A novel network, GPDCN, is proposed to address the structural distortions and enhance the precision of remote sensing imagery SR.
2) A unique DPDRB module is present for extracting multiscale feature maps without increasing the parameters via varied dilation rates. By this module, the entire pixel data can be covered and displayed.
3) A gradient propagation network (GPN) is designed to recover gradient maps from LR images to HR images, while additional supporting information is also provided for SR.
4) A GSA block is also developed to capture global contextual information and, simultaneously, to fully use self-similarity among nonadjacent pixels and improve the completeness of boundaries.

The rest of this article is organized as follows. Section II provides a synopsis of the relevant works. Section III will detail the proposed approach. Section IV discusses the experimental procedures and outcomes to validate the proposed method. Finally, Section V concludes this article.

II. Related Work

A. Edge Prior

The use of edge prior has been validated in previous work [11], which attempted to construct the gradient transfer mapping from LR to HR images using mathematical equations. Since an image's sharp edges correlate to well-defined gradients along the border, Fattal [11] tried using edge statistics to infer the prior reliance of various resolutions. Sun et al. [18] used a gradient field transformation to constrain the gradient field of the HR image and the reconstructed image. Furthermore, Tai et al. [19] extended edge-directed SR by using user-supplied example texture and restoring the fine details. Kondo and Fujiwara [20] combined reconstruction- and example-based SR to maintain natural edge structures. Yan et al. [21] enhanced the gradient profile sharpness, an edge sharpness metric, with a triangle model and a Gaussian mixture model.

Undoubtedly, it is unreliable to model the mapping relationships with a few parameters, especially with complicated land covers and intricate and fragmented image details. Therefore, the reconstructed images usually exhibit spurious artifacts or jagged edges. Since deep learning networks are good at end-to-end pixel transformations, several deep learning strategies have leveraged the advantage of prior's powerful qualities in SR assignments. Yang et al. [22] applied an off-the-shelf edge detector in a recurrent residual network to reconstruct fine features guided by edges. Ma et al. [23] introduced image gradient into the GAN-based SR network to provide additional structure information, improving the edge details' reconstruction ability.

B. Dilated Convolution

DC was first proposed to solve resolution reduction and information loss problems, which often appear in pixel-level tasks, such as image semantic segmentation. Early segmentation methods primarily conducted the pooling operation after the convolution layers to reduce the model's computation and increase the receptive field of the convolution layer. However, as the output of image segmentation is the pixel level, the output and input sizes should be the same. Thus, the deconvolution operation is commonly conducted in the latter part of the network, resulting in more missing information. Then, DC is introduced into semantic segmentation to solve this problem [24]. Subsequently, a few SR technologies have paid attention to the DC, as it is the same pixel-level task as semantic segmentation. Lin et al. [25] proposed a seven-layer dilated convolutional neural network with skip connections for reconstructing the HR image from an LR image. Mirchandani and Chordiya [26] presented a dilation patch super-resolution generative adversarial network by applying dilated operation in the generator architecture to obtain high-quality features. However, these methods have not focused on the massive potential of the dilation convolution to extract multiscale features. Different from these methods, this article exploits the different dilation rates to obtain multiscale features with a double-path extraction architecture.

C. Self-Attention Mechanisms

As a means of reallocating available resources to the most informative segments of inputs and modeling long-range dependencies, the attention mechanism was first applied in [27] to obtain global dependencies in the machine learning task. In addition, attention modules have been frequently used in CNNs for a variety of tasks, such as visual question answering [28] and image and video classification tasks [29], [30]. Attention modules are also becoming more widespread in the image SR field. Zhang et al. [31] applied SENet in CNNs to enhance SR performance. Dai et al. [32] constructed a second-order attention network (SAN) by considering the second-order characteristics of features. Fu et al. [33] presented dual attention as a technique for adaptively integrating local features with their global interdependence in both spatial and channel dimensions. Even though the attention mechanism has performed admirably in computer vision, when it comes to extracting explicit material from small and sparse elements, a lack of awareness of the relationship and correlation of each location is a big problem for image SR. Consequently, we have designed a GSA module to build the long-range correlation by integrating a query-specific global context for each query position.

D. SISR Algorithms of Sensing Images

SISR algorithms have gained popularity for remote sensing images in recent years [34]. Besides applying cutting-edge imaging technology, the SR technique is a low-cost and effective way of enhancing image quality. With the popularity of neural networks, numerous attempts have been made to design various architectures to obtain high-quality HR remote sensing images through learning a mapping function between LR and HR matches. Lei et al. [35] presented a local–global combination network (LGCNet) for image SR that was inspired by the success of CNN in natural image SR. To learn the connections between the characteristics from each recursion, Chang and Luo [36] devised a bidirectional convolutional long short-term memory layer. Jiang et al. [37] developed an edge-enhanced GAN model EEGAN for promoting satellite image SR reconstruction using an adversarial learning technique that can restore the sharp edge effectively. Dong et al. [38] proposed RRSGAN, which aligns the Ref features to the LR features, and the texture information in the Ref features can be transferred to the reconstructed HR images. To fuse multiscale high-/low-dimensional features, TransENet [39] introduced the transformer structure into the conventional SR framework and achieved superior performance in the remote sensing field. Zhang et al. [40] proposed a mixed high-order attention network (MHAN) to reconstruct the details by extracting high-order statistics. For hyperspectral imagery, Hang et al. [41] combined a decomposition subnetwork and a self-supervised subnetwork to construct an end-to-end SR network. Zhou et al. [42] proposed a pyramid fully convolutional network consisting of an encoder subnetwork and a pyramid fusion subnetwork to enhance the spatial resolution of low-spatial-resolution hyperspectral image.

III. Methodology

This section provides an overview of the framework. Unlike previous methods using sophisticated structures to form the deep architecture, the proposed model (see Fig. 1) is divided into two parts: a multiscale feature extraction network (the left part of Fig. 1) and a feature reconstruction network (the right part of Fig. 1). The feature extraction network is, in turn, divided into two parts: a structure-maintaining network (SN) and a GPN. The SN includes nine DPDRBs and five GSA modules. The GPN comprises three DPDRBs to obtain multiscale hierarchical features using different dilation rates. The feature reconstruction network is the MHOA, consisting of several different R-order ($R = 1, 2, 3, 4$) attention modules that reconstruct complicated details. The feature fusion and propagation from the extraction to the reconstruction are achieved through the frequency-based feature combination method.

A. Multiscale Feature Extraction Network

The multiscale feature extraction network includes an SN and a GPN and uses the three-band images as input. The SN begins with a convolution layer with a kernel size of 3×3 on the input image. Then, nine repeated DPDRBs follow in the later part of the network. After each of the first five layers, the feature maps are fed into the GSA block, gathering and distributing long-range image features. The DPDRB is inspired by multiscale residual block [43] and DC [24]. We incorporate the feature maps from the first, fifth, and ninth DPDRB to the GPN.

1) Double-Path Dilated Residual Block: Like image semantic segmentation, a pixel-level task, SR aims to predict an image's unknown pixels. Conducting multiple pooling operations would lose some vital feature information, resulting in unsatisfied reconstruction results. In addition, the standard 3×3 kernel focuses on a small region, ignoring nonadjacent pixels' relevance. Based on these considerations, we incorporate DC into the DPDRB (see Fig. 2), and the DC can be considered "convolution with a dilated filter," which is comparable to adding zero elements between two neighboring elements of the convolutional kernel. The dilation rate d means adding $d-1$ zero elements between the nearby elements of the kernel, and the receptive field by different dilation rates is presented in Fig. 3. By using variable dilation rates, the different receptive fields of each element can be achieved with the same convolutional kernel. In the DPDRB (see Fig. 2), we adopt a continually increasing dilation rate ($d=1, d=2$, and $d=3$) to accommodate exponentially expanding receptive fields.

The DPDRB can be illustrated as (1) and Fig. 2, which includes two parts: multiscale feature extraction and weighted residual connection. Given a feature F_{n-1} as the input of the DPDRB, S_1 and P_1 are the features through the first convolution layer, S_2 and P_2 are the features through the second convolution layer, and the final output feature F_n can be obtained as

$$S_1 = \sigma\left(w^1_{3\times3} \times F_{n-1} + b_1\right)$$
$$P_1 = \sigma\left(w^1_{d_{3\times3\_2}} \times F_{n-1} + b_1\right)$$
$$S_2 = \sigma\left(w^2_{3\times3} \times [S_1, P_1] + b_2\right)$$

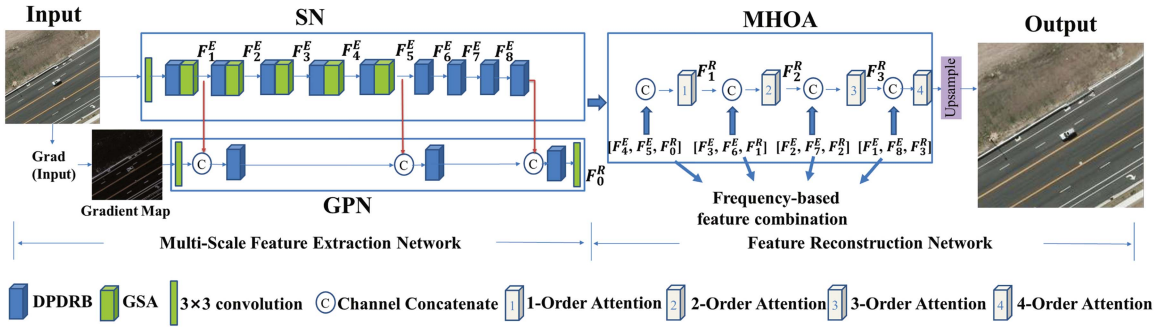


Fig. 1. Framework of the proposed GPDCN.

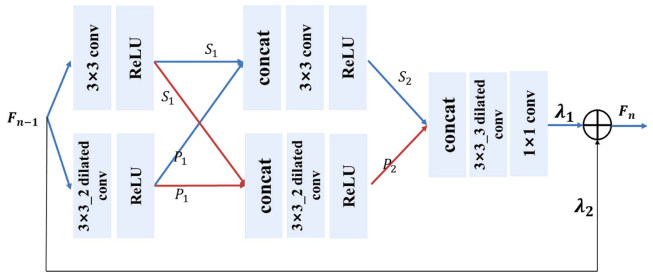


Fig. 2. Framework of the DPDRB.

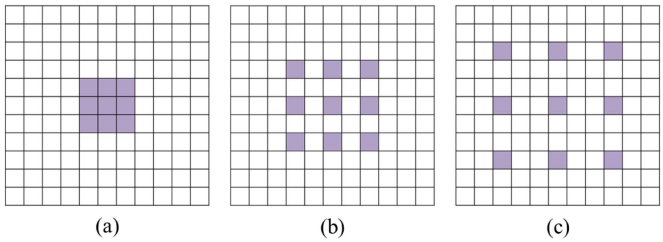


Fig. 3. Different receptive field for a 3×3 kernel by setting different dilation rate. (a) Dilation rate $= 1$ and receptive field $= 3 \times 3$. (b) Dilation rate $= 2$ and receptive field $= 5 \times 5$. (c) Dilation rate $= 3$ and receptive field $= 7 \times 7$.

$$
\begin{aligned}
P_2 &= \sigma\left(w^2_{d_{3\times3\_2}} \times [P_1, S_1] + b_2\right) \\
F &= w^3_{d_{1\times1}} \times \left(w^3_{d_{3\times3\_3}} \times [S_2, P_2] + b_3\right) \\
F_n &= \lambda_1 F + \lambda_2 F_{n-1} \qquad (1)
\end{aligned}
$$

where w and b denote the weights and bias, respectively, and the superscripts of w and the subscripts of b represent the layer index. The subscript 3×3 represents the size of the convolutional kernel. If the DC operation is used in the convolutional kernel, the subscript $d_{k\times k\_r}$ represents the $k \times k$ convolutional kernel with a dilation rate of r. $[S_1, P_1]$, $[P_1, S_1]$, and $[S_2, P_2]$ represent the concatenation operations, and λ_1 and λ_2 are the learnable parameters for weighting the input and output of the nonlinear mapping mode, respectively.

In the multiscale feature extraction part, a double-path DC network is adopted to extract local multiscale features by conducting different dilation rates. Based on this flexible DC, not only does the residual structure allow for bypassing abundant low-frequency information via several extraction layers, but also λ_1 and λ_2 act as learnable parameters during training to reallocate available resources toward the most informative components of the two parts, which encourage networks to focus on learning high-frequency information. Then, the embedding information can be shared in the block and, thus, transmit and fuse the image features of different scales.

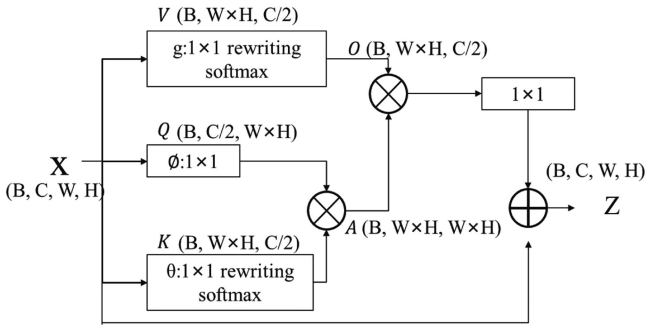


Fig. 4. Illustration of the GSA module.

2) GSA Module: The present convolutional network concentrates on the local region, suffering the limitation of capturing the global relationship among the whole spatial area. Considering the self-similarity in the long-range area in remote sensing images, the attention mechanism is widely used in recognition and detection tasks to aggregate the self-similar patches. Inspired by [30], [44], and [45], the GSA block is proposed for aggregating long-range feature information. The GSA block can be illustrated as Fig. 4, and the input feature maps can be defined as $X \in R^{B\times C\times W\times H}$, where B and C denote the batch size and channel number, respectively, and W and H represent the width and height of the input feature, respectively.

The input vectors are first fed into three 1×1 convolutions, and selective rewriting operations are performed to obtain three embedding feature maps K, Q, and V in different feature space, where $K \in R^{B\times(W\times H)\times(C/2)}$, $V \in R^{B\times(W\times H)\times(C/2)}$, and $Q \in R^{B\times(C/2)\times(W\times H)}$. To decrease the computation cost, we set the output channel as $C/2$ for dimension reduction. The attention weight matrix $A \in R^{B\times(W\times H)\times(W\times H)}$ can be obtained by calculating the self-similarity of K and Q as follows:

$$A = \text{softmax}(K) \otimes Q \qquad (2)$$

where $\otimes$ denotes matrix multiplications. Then, we rescale the value vector by conducting matrix multiplications on A and V, which can be illustrated as follows:

$$\begin{aligned} O &= A \otimes \text{softmax}(V) \\ &= [\text{softmax}(K) \otimes Q] \otimes \text{softmax}(V). \end{aligned} \tag{3}$$

Then, the convolution and elementwise addition operations yield the final output Z

$$Z = C_{1\times1}(O) + X \tag{4}$$

where C is the 1×1 convolution operation to adjust the channel number to the same as X.

In addition, (5) shows the GSA's matrix operations on a 3-D input array, ignoring the batch size B. In the first step, for the matrix K, K_i represents the values of all channels at the ith pixel position in space K, and Q_j denotes the values of all channels at the jth pixel location in space Q. $\otimes$ is the matrix multiplication operation. The element A_{ij} in attention matrix A can be viewed as the influence of the jth element of the input feature map on the ith element, thus realizing the dependence between any two elements of the global context. In the second step, we conduct the matrix multiplications on attention map A and value matrix V, where V_n represents the nth channel values for all positions. After conducting a softmax operation on matrix V, the matrix VS can be achieved. The multiplication processes can be described as aggregating the nonlocal self-similarity to achieve modeling of global long-range dependencies

$$A_{ij} = \text{softmax}(K_i \otimes Q_j) = \text{softmax}\left(\sum_{p=1}^{C/2} K_{ip} Q_{pj}\right)$$

$$A = \text{softmax}(K \otimes Q)$$

$$Z_{mn} = A_m \otimes [\text{softmax}(V)]_n = A_m \otimes VS_n = \sum_{q=1}^{W\times H} A_{mq} VS_{qn}$$

$$\begin{aligned} Z &= A \otimes \text{softmax}(V) \\ &= [\text{softmax}(K \otimes Q)] \otimes \text{softmax}(V). \end{aligned} \tag{5}$$

GSA blocks enhance the description of pixel features by modeling long-range relationships. Optimized for the first attention mechanism Nonlocal block [30], the GSA performs two softmax operations to smooth the image and neutralizes the effect of the sharp gradient map.

3) Gradient Propagation Network: The gradient of the image denotes the change rate of the image pixels' gray value along the x-axis and y-axis. The contours of HR images are sharper than those of LR images, as shown in Fig. 5. In other words, a more excellent gradient value indicates a clearer margin. As an image can be viewed as a 2-D discrete function, the difference of the neighboring pixels can be approximately calculated as the gradient

$$\begin{aligned} G_x(x,y) &= I(x+1,y) - I(x-1,y) \\ G_y(x,y) &= I(x,y+1) - I(x,y-1) \\ \nabla G(x) &= (G_x(x,y), G_y(x,y)) \end{aligned}$$

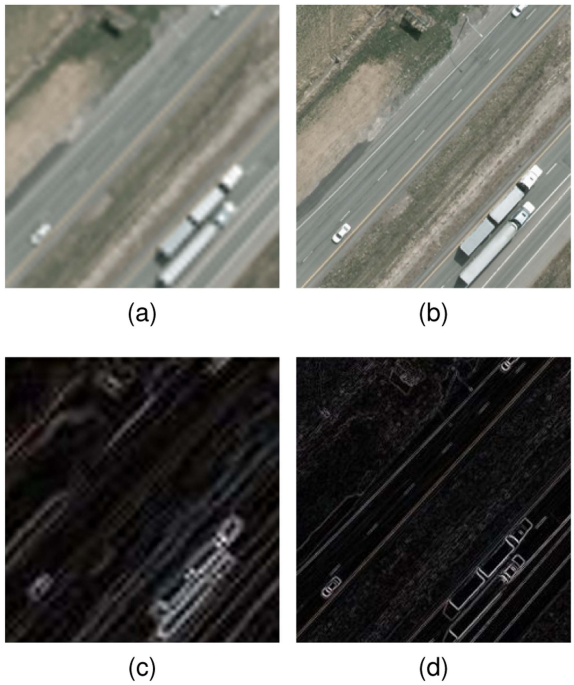


Fig. 5. Gradient difference of the LR image and HR image. (a) LR image. (b) HR image. (c) Gradient of the LR image. (d) Gradient of the HR image.

$$\text{Grad}(G) = \|\nabla G\|_2 = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{6}$$

where $I(x,y)$ represents the pixel value at the (x,y) position; $G_x(\cdot)$ and $G_y(\cdot)$ represent the gradient value along the x and y axes, respectively, and a convolutional kernel can achieve it; $I_x(\cdot)$ stands for the pixel value of the according coordinates; and $\text{Grad}(\cdot)$ is the gradient map at coordinate (x,y).

As shown in Fig. 5, the gray value changes sharply on edge, so the gradient map can illustrate the sharpness feature of the image. In addition, this valuable feature can be utilized to facilitate the sharpness of the SR images. Considering this, the GPN is proposed to propagate the gradient in the network. The GPN takes the gradient map $\text{Grad}(G)$ obtained from the gradient calculation as the input and starts with an initial convolution layer with a kernel size of 3×3. In the latter of the GPN, there are three repeated gradient blocks (GBs), each of which is the same as the DPDRB, concatenating the intermediate-level SR feature and gradient feature as the input

$$f_{\text{input}} = [F_{\text{sr}}, F_{\text{gradient}}] \tag{7}$$

where F_{sr} and F_{gradient} represent the intermediate-level SR feature and gradient features, respectively. Since a well-designed SR branch can convey abundant edge structural information, the proposed strategy performs outstandingly for recovering gradient maps. Simultaneously, the feature maps are also vital and are employed prior to enhancing the GPN's performance. With three propagating GBs, the integrated feature maps are extracted and fed into the last convolution layer with a 3×3 kernel. The final output of the GPN is conducted as part of the input for the reconstruction network.

B. Feature Reconstruction Network

1) Mixed High-Order Attention Module: Attention adjusts the weight of the convolution parameters to highlight the essential parts and reduce the influence of useless information. The traditional channel-based [31] and spatial-based [46] attention commonly used in the CNN concentrates on the first-order attention, which only obtains the coarse feature statistics because they only focus on a specific region, failing to accurately predict the impact of different regions operating in concert on the final result. To model the complicated and high-order feature statistics for complex details in remote sensing images and model the attention of different parts of the image or feature map acting in concert mechanism, we introduce the HOA module [40], [47] into the reconstruction network, in which R is defined as the order of the HOA; more specifically, when $R > 1$, the R-order attention concentrates on R regions and models the interaction of R attentions, which can have a synergistic effect on the final output.

For an HOA module with R order, we use R 1×1 convolutions to obtain R embedding feature maps $\{Z_s^R\}_{s=1,\ldots,R}$ at level R. At level $R-1$, we also use $R-1$ convolutions to obtain $R-1$ embedding feature maps $\{Z_s^{R-1}\}_{s=1,\ldots,R-1}$. Until level 1, $R(R+1)/2$ feature maps are achieved. In the feature map sets $\{Z_s^r\}_{s=1,\ldots,r}$, where $r = 1, 2, \ldots, R$, we combine the R-order feature statistic by means of the elementwise product, which can be formulated as $Z^r = Z_1^r \odot Z_2^r \odot \cdots \odot Z_r^r = \prod_{i=1}^{r} Z_i^r$.

In addition, we apply the nonlinearity variation to improve the representation capacity of this high-order feature map, and the variation is formulated by

$$A(X) = \text{sigmoid}\left(\sum_{r=1}^{R} \text{ReLU}\left(Z^r\right)\right) \tag{8}$$

where the ReLU function denotes the nonlinear activation function, and the sigmoid function is applied to restrict the element of the $A(X)$ in [0,1]. Finally, similar to the general attention mechanism, $A(X)$ is used to reweight the input X, that is, $Y = A(X) \odot X$.

More specifically, as shown in Fig. 6, when $R = 1$, we get a descriptor Z, and $A(X) = \text{sigmoid}(Z)$, then $Y = A(X) \odot X$, as shown Fig. 6(a); when $R = 2$, we first use two 1×1 convolutions to get two embedding features, $\{Z_1^2, Z_2^2\}$, at level 2. We combine them to achieve the second-order component, that is, $Z^2 = Z_1^2 \odot Z_2^2$. Then, at level 1, one convolution is used to obtain an embedding feature, Z^1. Subsequently, the weighted matrix is calculated as $A(X) = \text{sigmoid}(\text{ReLU}(Z^1) + \text{ReLU}(Z^2))$, and finally, the output Y can be obtained by the operation $Y = A(X) \odot X$, as shown in Fig. 6(b).

2) Frequency-Based Feature Combination: In the present SR methods, many residual blocks are usually accumulated to model a very deep network, and the feature map from the last residual block is then commonly used to reconstruct the HR images. Though these "deep" networks have achieved some satisfying performance, there is a limitation in considering the feature distribution of hierarchical features, failing to utilize the feature map of different frequencies. Qiu et al. [48] pointed out that the feature maps from different frequency bands often vary greatly. For shallow layers, the parameters concentrate on low-frequency information, such as the basic textures. For deeper layers, the parameters emphasize high-frequency components of the whole image, including the region filled with edges, complicated corners, and other characteristics. As a result, low-frequency feature maps from shallow layers should be routed to a higher order HOA module, which is more elaborate and has a greater capacity for detail restoration. In addition, high-frequency components are necessary for reconstruction, and the information collected at higher frequencies from deep layers should be augmented further using a higher order HOA module. Thus, it is not the best way to concatenate all the feature maps obtained directly from different layers. Therefore, the feature maps from shallow layers and deep layers are combined based on various frequencies in our reconstruction network, and the process can be formulated as

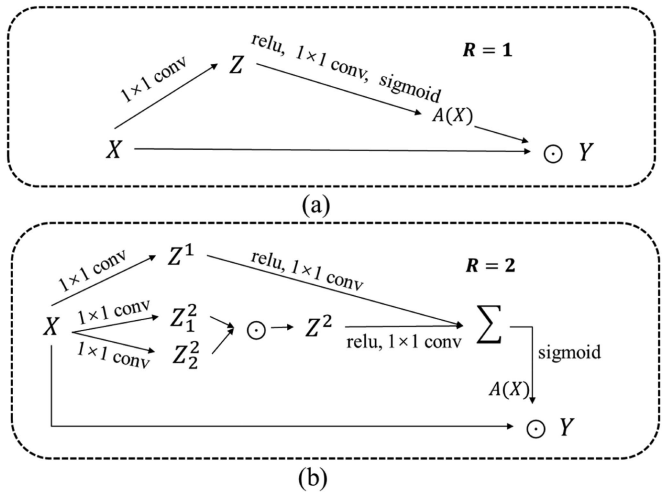


Fig. 6. Illustration of HOA structure with (a) $R = 1$ and (b) $R = 2$.

$$F_r^R = \text{HOA}_r([F_{N-r+1}^E, F_{M-(N+r-1)}^E, F_{r-1}^R]). \tag{9}$$

Assume that the reconstruction network contains N HOA modules and $M = 2N + 1$. F^E and F^R denote the feature maps from the extraction network and the reconstruction network, respectively, $[\cdot]$ is the concatenation conduction, HOA_r denotes the HOA module of order r, and F_r^R and F_{r-1}^R, respectively, represent the output feature from the r-order HOA module and the $(r-1)$-order HOA module in the reconstruction part.

C. Loss Function

The proposed method, GPDCN, aims to minimize the difference between the reconstructed and ground-truth images. Considering that the majority of image evaluation indicators are significantly connected to pixel-by-pixel differences, pixel loss is still highly desirable. Based on the pixel loss between the HR image I_{HR} and the SR image I_{SR}, we add an additional gradient loss to consider the gradient information. The loss function in our experiment consists of two parts: image pixel loss and

gradient loss

$$\begin{aligned}\mathcal{L}_{\text{image}} &= L_1\left(I_{\text{SR}}, I_{\text{HR}}\right) = \frac{1}{hwc}\sum_{i,j,k}\left|I_{\text{SR}}^{i,j,k} - I_{\text{HR}}^{i,j,k}\right| \\ \mathcal{L}_{\text{gradient}} &= L_1\left(I_{\text{SR\_grad}}, I_{\text{HR\_grad}}\right) \\ &= \frac{1}{hwc}\sum_{i,j,k}\left|I_{\text{SR\_grad}}^{i,j,k} - I_{\text{HR\_grad}}^{i,j,k}\right| \\ \mathcal{L}_{\text{total}} &= L_{\text{image}} + \lambda L_{\text{gradient}} \end{aligned} \tag{10}$$

where h, w, and c represent the height, width, and channel number of the picture, respectively, and λ is the weighted parameter on the gradient loss. In this article, we set $\lambda = 0.001$.

IV. Experiments and Analyses

A. Datasets and Implementation Details

1) Datasets: The proposed approach is evaluated using a new SR dataset named 3K VEHICLE_SR, which is obtained from the 3K VEHICLE dataset, a standard vehicle identification dataset [49] consisting of 20 pictures with a spatial resolution of 0.13 m of 5616 × 3744 pixels. The 3K VEHICLE dataset includes vehicles in a variety of real-world environments, including ports, hills, lakes, metropolitan zones, and rivers. The 20 photographs are cropped into 1170 subimages of 512 × 512 pixels and are used in the experiment as 3K VEHICLE_SR (the sizes of the LR images are 128×128 for ×4 scale, 171×171 for ×3 scale, and 256×256 for ×2 scale). Eighty percent of the subimages are used for training, 10% for validation, and the remaining 10% for testing. In addition, the Massachusetts_Roads dataset is used to assess the generalizability and robustness of the technique. We also use the WHU-RS19 dataset to test the performances on different scenes. In the experiments, all the LR images are degraded from HR images by bicubic interpolation, and the corresponding HR ones are reviewed as ground truth.

2) Implementation Details: In the experiments, we focus on scale factors of ×2, ×3, and ×4 and evaluate SR results using the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) on the Y channel of the converted YCbCr space. The model inputs and outputs three-channel RGB pictures. Since L1 loss accelerates convergence and stabilizes the convergence process, the L1 loss is adopted as the training loss function. ADAM [50] is used to optimize the training process with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to update the model parameters. The initial learning rate is set at 10^{-4} and then drops by a factor of 10 every 100 epochs. All the experiments are conducted with PyTorch on NVIDIA TITAN RTX GPU.

B. Evaluation Metrics

The PSNR and SSIM, which are frequently used as assessment metrics for SR tasks, are employed in this experiment to quantify the suggested method. Given an HR image H and an SR image S, N denotes the image's total pixel number. The PSNR and SSIM can be obtained as follows:

$$\text{PSNR}(H,S) = 10\cdot\log_{10}\left(\frac{\text{MAXI}^2}{\frac{1}{N}\sum_{t=1}^{N}(H(t)-S(t))^2}\right) \tag{11}$$

where t represents the tth location, $H(t)$ and $S(t)$ are the corresponding pixel value of the HR and SR images, respectively, and MAXI equals 255 for the 8-bit image

$$\text{SSIM}(H,S) = \frac{(2u_Hu_S + c_1)(2\sigma_{HS} + c_2)}{(u_H^2 + u_S^2 + c_1)(\sigma_H^2 + \sigma_S^2 + c_2)} \tag{12}$$

where u is the mean pixel value, σ represents the pixel variance value, and σ_{HS} represents the covariance between the HR image and the SR image.

TABLE I
Ablation Study With Different Components' Combinations on the 3K VEHICLE_SR Dataset

DC	×	✓	✓	✓	×	✓
GSA	×	✓	✓	×	✓	✓
GPN	×	✓	×	✓	✓	✓
MHOA	×	×	✓	✓	✓	✓
PSNR	29.4582	29.6882	29.6931	29.6793	29.7365	**29.7775**
SSIM	0.6981	0.7062	0.7068	0.7066	0.7087	**0.7094**

C. Ablation Study

Since our GPDCN contains a trick DC and three major components—GSA module, GPN, and MHOA network—to demonstrate the validity of different parts, five tests are designed and conducted. The first column represents baseline, and the second to fifth columns represent the results of removing a module by the model, in order to verify the effectiveness of the modules. The results in Table I demonstrate that the GSA module contributes more improvement than the GPN and MHOA do. The GSA module, GPN, and MHOA can increase the network's performance for the results of ×4 SR by 0.0982, 0.0844, and 0.0893 dB, respectively. Obviously, when all the components are integrated, further promotion is achieved, demonstrating the importance of each design in obtaining the highest SR results in the proposed network.

D. Comparisons With State-of-the-Art Methods

To demonstrate the feasibility of our method, we compare it with other representative interpolation-, CNN-, and ResNet-based SR approaches: ESPCN [51], Bicubic [9], VDSR [52], DRN [53], IGNN [54], RCAN [31], SAN [32], and RDN [55]. We analyze various comparison approaches using open-source code, and all the methods are trained and tested in the same environment to guarantee the results are credible.

1) Quantitative Evaluations: Table II details the PSNR and SSIM values. Table II shows that the GPDCN outperforms the current competitive approaches on two metrics (PSNR and SSIM) for almost all factors except for the ×2, and the reason may be that the GSA's advantages of catching self-similar properties are not particularly obvious for the images with relatively

TABLE II
QUANTITATIVE COMPARISON ON MASSACHUSETTS_ROADS AND 3K VEHICLE_SR USING DIFFERENT METHODS

Method	scale	Massachusetts_Roads		3K VEHICLE_SR	
		PSNR	**SSIM**	**PSNR**	**SSIM**
ESPCN	×4	22.073	0.573	27.422	0.649
Bicubic		22.936	0.516	28.831	0.619
VDSR		23.306	0.581	29.371	0.697
DRN		22.179	0.526	28.381	0.663
IGNN		23.717	0.609	29.405	0.697
RCAN		23.819	0.613	29.748	**0.709**
SAN		23.743	0.607	29.75	0.708
RDN		23.848	**0.616**	29.745	**0.709**
GPDCN(ours)		**23.865**	**0.616**	**29.778**	**0.709**
ESPCN	×3	22.907	0.647	27.732	0.692
Bicubic		23.868	0.616	29.305	0.687
VDSR		23.867	0.647	29.463	0.725
DRN		22.752	0.574	27.903	0.669
IGNN		24.311	0.658	29.662	0.739
RCAN		24.322	0.666	29.916	0.74
SAN		24.313	0.666	29.902	0.739
RDN		24.281	0.668	30.008	0.744
GPDCN (ours)		**24.491**	**0.678**	**30.032**	**0.745**
ESPCN	×2	24.816	0.778	30.421	0.826
Bicubic		25.684	0.75	31.713	0.792
VDSR		26.872	0.778	32.925	0.823
DRN		25.316	0.752	31.18	0.823
IGNN		**26.924**	0.816	33.103	0.859
RCAN		26.805	0.815	33.069	0.859
SAN		26.821	**0.817**	33.154	**0.861**
RDN		26.706	0.816	**33.183**	**0.861**
GPDCN (ours)		26.867	**0.817**	33.12	0.86

The values in bold are the best.

clear contours. When the amplification factor is set to 4, our GPDCN model achieves the highest PSNR across both the testing datasets. In the 3K VEHICLE_SR dataset, the proposed GPDCN outperforms Bicubic, VDSR, SAN, and RDN by approximately 0.947, 0.407, 0.028, and 0.033 dB, respectively. For the Massachusetts_Roads dataset, our model outperformed the SAN by 0.122 dB, a classical attention-based network that investigates the second-order statistics of feature maps, and the results validate the effectiveness of high-order characteristics. In addition, compared with the state-of-the-art approach, IGNN [54], a cross-scale nonlocal SR network, the proposed algorithm outperforms it by 0.148 dB. The Massachusetts_Roads dataset is densely packed with repetitive patterns, such as edges and small corners. As a result, the GPDCN achieves superior performance and demonstrates the efficacy of the proposed method at restoring edges and small details, which indicates

TABLE III
COMPARISON OF THE RUNNING TIME, PARAMETERS, FLOPS, GPU COST, AND PSNR ON TWO DATASETS WITH THE SCALE FACTOR OF 4

Method	Time (s)	Number of parameters	FLOPs (GFLOPs)	GPU (M)	**PSNR**	
					3K VEHICLE_SR	Massachusetts Roads
VDSR	0.079	668K	10.98	163	29.283	23.605
DRN	4.371	5M	205.01	1452	28.381	22.179
IGNN	0.305	44.8M	1044	566	29.405	23.717
LGCNet	0.039	193K	3.19	53	29.275	23.274
RCAN	0.215	16M	261.67	1059	29.748	23.819
SAN	0.224	15.7M	91.62	489	29.75	23.743
RDN	0.062	22.3M	278.48	1219	29.745	23.848
GPDCN	0.079	14.7M	202.78	338	**29.778**	**23.865**

that gradient priors and global context-aware information are powerful tools for a more accurate restoration.

2) Qualitative Evaluations: We also examine the visual effects of various procedures, as depicted in Fig. 7. We highlight the regions of interest that clearly illustrate the distinctions among the different methods. The results reveal that the fundamental bicubic interpolation technique cannot increase the amount of information. Deep-learning-based methods, such as VDSR, are capable of inferring some texture information but produce blurry image contours as a result of their global optimization method and wasteful feature usage. The results generated by the proposed approach are very competitive and much more realistic for images containing repeating high-frequency characteristics, such as corners, lines, and squares. For instance, in the figure of a car [see Fig. 7(e)], only the proposed method can reestablish an accurate and evident pattern with fewer artifacts, whereas others suffer from different degrees of blurring. Fig. 8 illustrates the gradient maps. As we can see, other methods' gradient maps have low values or include structure degradation, whereas ours are bold and realistic. The qualitative comparison demonstrates that our proposed GPDCN method is capable of extracting additional structure information from the gradient space to generate clear and realistic SR images by maintaining geometric structures.

3) Model Size Analysis: Since the model size is an essential consideration in practical applications, particularly on limited computer devices, we provide the model size and performance for currently competitive SR approaches in Table III. According to Table III, compared with DRN [53] and SAN [32], although the proposed method processes more parameters, the PSNR value is still much enhanced. Furthermore, compared with IGNN [54] and RDN [55], our algorithm produces more competitive results with fewer parameters. At the same time, the FLOPs and GPU costs of our method are comparable. As illustrated in Table III, our approach balances effectiveness and efficiency.

4) Comparison With Remote Sensing SR Methods: GPDCN's advantages in various aspects, including model size and memory cost, have been proved in previous experiments. However, all the comparison approaches used in the preceding investigations are proposed for natural image SR. To further validate the GPDCN's performance, we compare it with several domain-specific SR methods. In this part, five approaches to

Fig. 7. Reconstruction visual results on 3K VEHICLE_SR and Massachusetts_Roads on a scale of $\times 4$ for different SR methods. (a)–(d) are from the 3K VEHICLE_SR dataset.

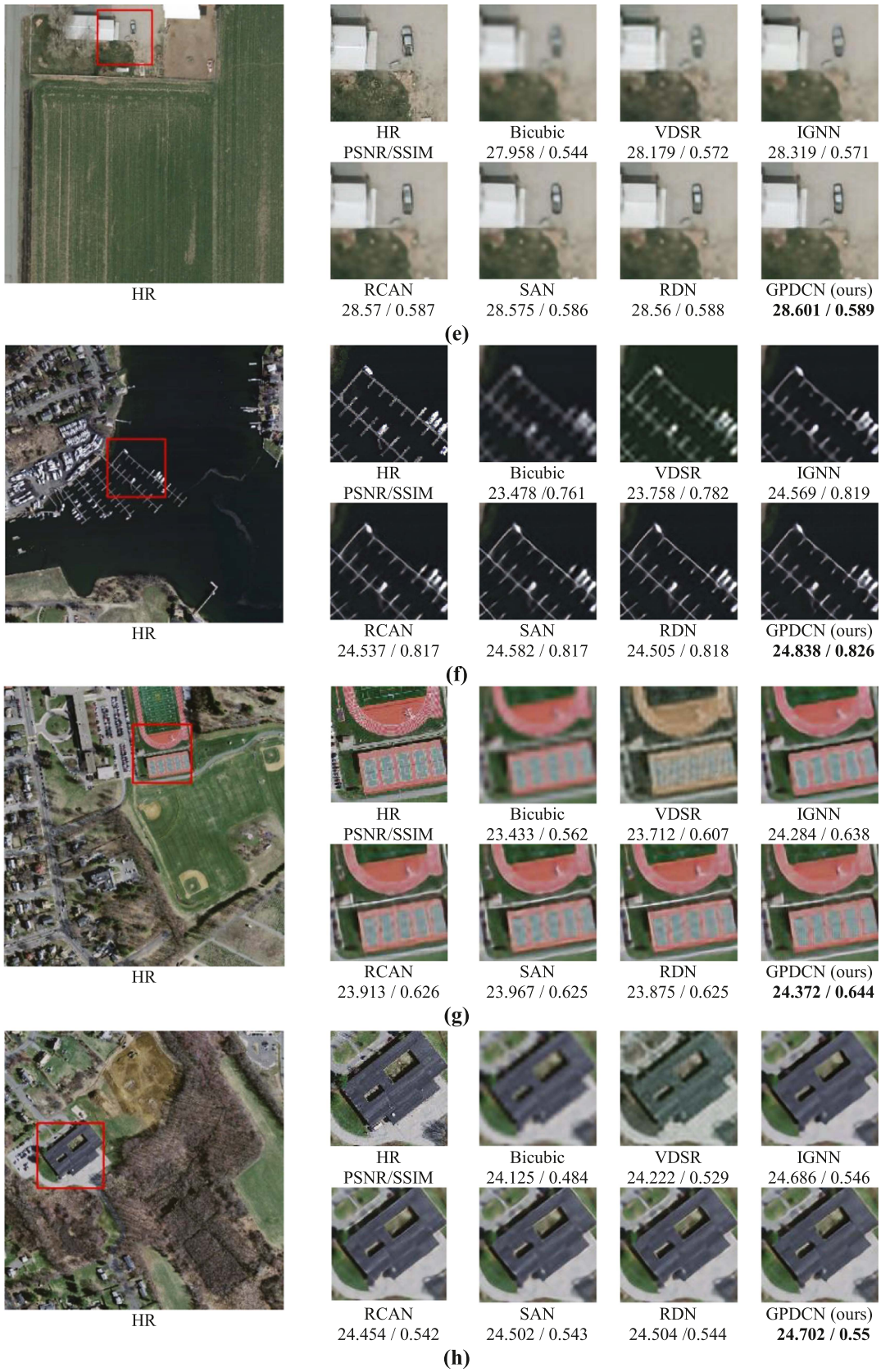


Fig. 7. (Continued.) Reconstruction visual results on 3K VEHICLE_SR and Massachusetts_Roads on a scale of ×4 for different SR methods. (e) is from the 3K VEHICLE_SR dataset. (f)–(h) are from the Massachusetts_Roads dataset.

Fig. 8. Visual results of gradient maps with state-of-the-art SR methods. The proposed GPDCN method can better maintain gradients and structures.

TABLE IV
COMPARISON RESULTS OF PSNR, SSIM, NIQE, PI, AND LPIPS ON THE 3K VEHICLE_SR DATASET WITH A SCALE FACTOR OF 4

Method	PSNR	SSIM	NIQE	PI	LPIPS
EEGAN	29.116	0.687	7.044	6.522	0.596
LGCNet	29.275	0.694	6.722	6.134	0.574
RDBPN	29.705	0.706	**6.588**	6.071	0.54
TransENet	29.727	0.708	7.081	6.247	0.547
MHAN	29.715	0.708	6.595	6.065	0.545
GPDCN	**29.778**	**0.709**	6.978	**5.856**	**0.44**

The best scores are highlighted.

remote sensing images SR are discussed in detail: EEGAN [37], LGCNet [35], RDBPN [56], TransEnet [39], and MHAN [40]. Except for PSNR and SSIM, Natural Image Quality Evaluator (NIQE) [57], Perceptual Index (PI) [58], and learned perceptual image patch similarity (LPIPS) [59] are introduced into the evaluation as PIs. Note that the lower the NIQE, PI, and LPIPS, the higher the quality of the images. The conclusions are shown in Table IV, and it is clear that our approach performs the best on the 3K VEHICLE_SR dataset with a scale factor of $\times 4$. In addition, we present a visual comparison in Fig. 9. By zooming the regions of interest, it is evident that the EEGAN and LGCNet cannot produce rich details. RDBPN and TransENet can generate sharp edges but with some blurred contents. As for the MHAN, it can slightly improve the reconstruction performance by using the high-order attention mechanism while still having some noises and artifacts. And our GPDCN reconstructs the most realistic image features with the slightest ambiguity, resulting in more visually attractive results. It unequivocally establishes that the proposed GPDCN is a viable solution to the remote sensing imagery SR problem.

E. Comparison on Image Scene Classification Dataset

We compare our GPDCN with several SISR methods on WHU-RS19, a scene classification dataset for remote sensing image. Table V lists the detailed performances of different methods for scale $\times 4$ on all 19 classes. From the results, it can be observed that our model obtains the best PSNR results

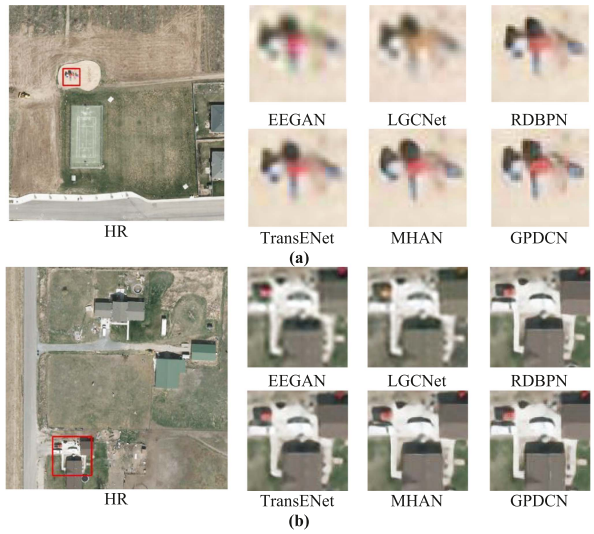


Fig. 9. (a) and (b) Visual results of different remote sensing SR-based methods.

in ten scene categories and the best SSIM results in 17 scene categories. Compared with other algorithms, our method is more effective in some scenes, which include abundant edges and contours, such as "Airport," "Bridge," "Commercial," "Park," etc. Meanwhile, the proposed method achieve best PSNR and SSIM for the overall evaluation. It also shows that the PSNR results are very different among different scenes, in which the PSNR for "Beach" images is 40.198 dB, but the PSNR for "Residential" images is only 22.418 dB. The reason for this is that the image includes a wide range of remote sensing scenes, i.e., the image of "Residential" owns more high-frequency information than the "Beach." The very smoothing scenes, where minimal high-frequency information should be super-resolved, tend to have higher PSNR results.

F. Visualization of Image Gradient

To demonstrate the efficacy of the GPN, the gradient maps as outputs are shown in Fig. 10. Given sharp-edged HR photos, the generated corresponding gradient maps often contain fine-grained and unambiguous contours for the items in the images

TABLE V
COMPARISON RESULTS OF PSNR AND SSIM ON THE WHU-RS19 DATASET WITH A SCALE FACTOR OF 4

WHU-RS19	RCAN		SAN		RDN		VDSR		RDBPN		EEGAN		GPDCN	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Airport	26.921	0.703	26.959	0.703	26.977	0.704	26.959	0.708	26.992	0.704	26.903	0.700	**27.043**	**0.708**
Beach	**40.307**	0.961	36.883	0.923	36.884	0.923	34.838	0.920	36.851	0.920	36.150	0.918	40.198	**0.963**
Bridge	32.642	0.861	32.749	0.862	32.721	0.861	30.140	0.858	32.711	0.856	32.398	0.855	**32.768**	**0.863**
Commercial	22.649	0.605	22.720	0.608	22.756	0.610	22.731	0.609	22.759	0.608	22.646	0.592	**22.797**	**0.612**
Desert	37.28	0.887	**37.326**	**0.888**	37.160	0.884	34.792	0.881	37.236	0.885	34.664	0.883	36.747	0.887
Farmland	34.896	0.827	34.851	0.826	34.846	0.825	33.418	0.828	**34.934**	**0.828**	34.551	0.824	34.868	0.826
footballField	26.875	0.710	26.891	0.711	26.918	0.711	26.414	0.709	**26.963**	0.711	26.372	0.691	26.947	**0.714**
Forest	26.046	0.512	26.094	0.514	26.045	0.512	26.000	0.511	26.030	0.510	25.977	0.505	**26.128**	**0.517**
Industrial	24.236	0.632	24.289	0.634	24.292	0.634	23.646	0.631	**24.332**	0.635	23.905	0.616	24.325	**0.637**
Meadow	34.294	0.783	**34.313**	0.784	34.219	0.783	30.378	0.781	34.245	0.781	33.139	0.769	34.134	**0.784**
Mountain	24.444	0.497	24.446	0.498	24.460	0.493	**24.517**	0.497	24.379	0.495	24.408	0.499	24.498	**0.500**
Park	26.887	0.649	26.883	0.649	26.921	0.650	26.480	0.651	26.909	0.648	26.842	0.650	**26.949**	**0.652**
Parking	25.701	0.718	25.733	0.718	25.731	0.718	25.870	0.720	**25.857**	0.722	25.646	0.708	25.796	**0.723**
Pond	30.016	0.805	30.062	0.806	30.074	0.806	29.355	0.807	30.050	0.804	29.665	0.777	**30.075**	**0.807**
Port	23.957	0.675	24.003	0.676	24.012	0.677	24.031	0.680	**24.096**	0.677	23.942	0.661	24.042	**0.679**
railwayStation	23.838	0.566	23.854	0.567	23.863	0.567	23.413	0.564	23.861	0.566	23.581	0.554	**23.948**	**0.571**
Residential	22.179	0.617	22.292	0.621	22.326	0.623	22.397	0.621	22.342	0.623	22.166	0.601	**22.418**	**0.629**
River	26.946	0.651	26.989	0.652	26.984	0.652	26.905	0.652	26.951	0.650	26.900	0.640	**27.048**	**0.655**
Viaduct	24.847	0.637	24.902	0.640	24.923	0.640	24.928	0.639	24.923	0.639	24.704	0.624	**25.013**	**0.645**
Average	28.156	0.700	28.013	0.699	28.006	0.698	27.222	0.698	28.022	0.698	27.608	0.688	**28.197**	**0.704**

The best scores are highlighted.

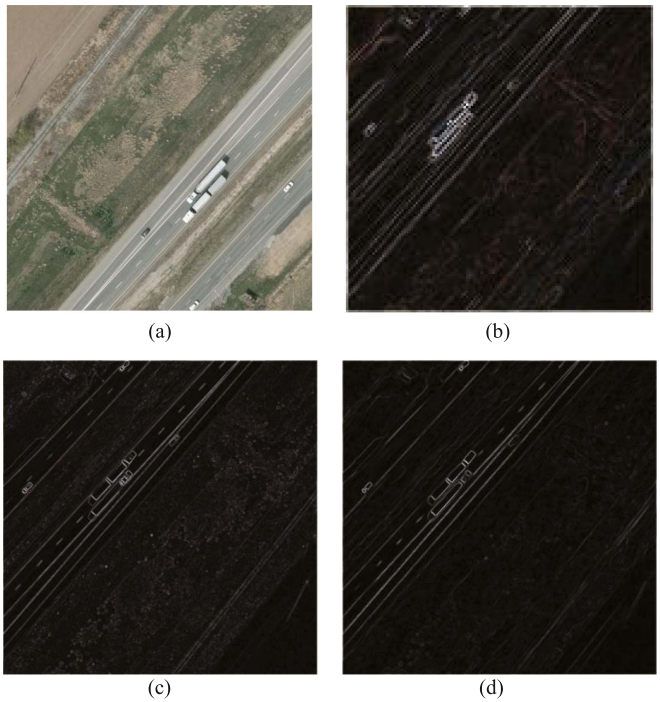


Fig. 10. Visualization of gradient maps ("00001093_co.png" from 3K VEHICLE_SR). (a) HR Image. (b) LR image gradient. (c) HR image gradient. (d) Output of the gradient network.

[see Fig. 10(c)]. As Fig. 10 shows, after the bicubic operation, the gradient maps generated from the LR equivalents frequently exhibit coarse-grained shapes [see Fig. 10(b)]. Our GPN inputs LR gradient maps and outputs SR gradient maps to supply the SR branch with precise structural information [see Fig. 10(d)]. The result shows that our GPN successfully recovers sharp and structure-pleasing gradient maps.

V. CONCLUSION

This article presented a GPDCN for SR in remote sensing images. To be more precise, the DPDRB was proposed to enlarge the receptive field and extract multiscale feature maps. The GSA structure enabled the GPDCN to acquire long-distance interactions and structural information to integrate nonlocal operations into the network. Meanwhile, we constructed a GPN using double-path DC blocks to recover HR gradient maps from the LR ones and provide explicit structural guidance to the SR branch via gradient information. In addition, an MHOA module was adopted to reconstruct the image using hierarchical characteristics with various frequency bands. Extensive experimental results demonstrated that the proposed GPDCN can surpass existing SR algorithms and balance performance and efficacy.

ACKNOWLEDGMENT

The authors would like to thank editors, associate editors, and anonymous reviewers for their insightful suggestions and comments, which significantly improve this article.

REFERENCES

[1] Z. Liu, G. Li, G. Mercier, Y. He, and Q. Pan, "Change detection in heterogenous remote sensing images via homogeneous pixel transformation," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1822–1834, Apr. 2018.
[2] J. Peng, L. Li, and Y. Y. Tang, "Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1790–1802, Jun. 2019.

[3] J. Peng, W. Sun, and Q. Du, "Self-paced joint sparse representation for the classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1183–1194, Feb. 2019.

[4] M. Thornton, P. Atkinson, and D. Holland, "A linearised pixel-swapping method for mapping rural linear land cover features from fine spatial resolution remotely sensed imagery," *Comput. Geosci.*, vol. 33, no. 10, pp. 1261–1272, 2007.

[5] X. Liu, D. Zhai, R. Chen, X. Ji, D. Zhao, and W. Gao, "Depth restoration from RGB-D data via joint adaptive regularization and thresholding on manifolds," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1068–1079, Mar. 2019.

[6] J. Gu, X. Sun, Y. Zhang, K. Fu, and L. Wang, "Deep residual squeeze and excitation network for remote sensing image super-resolution," *Remote Sens.*, vol. 11, no. 15, 2019, Art. no. 1817.

[7] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1122–1131.

[8] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Trans. Multimedia*, vol. 21, pp. 3106–3121, 2019.

[9] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 6, pp. 1153–1160, Dec. 1981.

[10] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[11] R Fattal, "Image upsampling via imposed edge statistics," *ACM Trans. Graph.*, vol. 26, no. 3, p. 95, 2007.

[12] S. Dai, M. Han, W. Xu, Y. Wu, Y. Gong, and A. K. Katsaggelos, "SoftCuts: A soft edge smoothness prior for color image super-resolution," *IEEE Trans. Image Process.*, vol. 18, no. 5, pp. 969–981, May 2009.

[13] W. Freeman, T. Jones, and E. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, Mar./Apr. 2002.

[14] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3791–3799.

[15] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1920–1927.

[16] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[17] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1637–1645.

[18] J. Sun, Z. Xu, and H.-Y. Shum, "Gradient profile prior and its applications in image super-resolution and enhancement," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1529–1542, Jun. 2011.

[19] Y.-W. Tai, S. Liu, M. S. Brown, and S. Lin, "Super resolution using edge prior and single image detail synthesis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2400–2407.

[20] K. Kondo and H. Fujiwara, "Edge preserving super-resolution with details based on similar texture synthesis," in *Proc. IEEE Asia Pacific Conf. Circuits Syst.*, 2014, pp. 29–32.

[21] Q. Yan, Y. Xu, X. Yang, and T. Q. Nguyen, "Single image superresolution based on gradient profile sharpness," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3187–3202, Oct. 2015.

[22] W. Yang et al., "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5895–5907, Dec. 2017.

[23] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-preserving super resolution with gradient guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7769–7778.

[24] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2016, *arXiv:1511.07122*.

[25] G. Lin, Q. Wu, L. Qiu, and X. Huang, "Image super-resolution using a dilated convolutional neural network," *Neurocomputing*, vol. 275, pp. 1219–1230, 2018.

[26] K. Mirchandani and K. Chordiya, "DPSRGAN: Dilation patch super-resolution generative adversarial networks," in *Proc. 6th Int. Conf. Convergence Technol.*, 2021, pp. 1–7.

[27] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5197–5206.

[28] J. Chen, J. Shao, and C. He, "Movie fill in the blank by joint learning from video and text with adaptive temporal attention," *Pattern Recognit. Lett.*, vol. 132, pp. 62–68, 2020.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[30] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[31] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.

[32] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11065–11074.

[33] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[34] D. Wu, Y. Zhang, Y. Chen, and S. Zhong, "Vehicle detection in high-resolution images using superpixel segmentation and CNN iteration strategy," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 105–109, Jan. 2019.

[35] S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1243–1247, Aug. 2017.

[36] Y. Chang and B. Luo, "Bidirectional convolutional LSTM neural network for remote sensing image super-resolution," *Remote Sens.*, vol. 11, no. 20, 2019, Art. no. 2333.

[37] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.

[38] R. Dong, L. Zhang, and H. Fu, "RRSGAN: Reference-based super-resolution for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5601117.

[39] S. Lei, Z. Shi, and W. Mo, "Transformer-based multi-stage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5615611.

[40] D. Zhang, J. Shao, X. Li, and H. T. Shen, "Remote sensing image super-resolution via mixed high-order attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5183–5196, Jun. 2021.

[41] R. Hang, Q. Liu, and Z. Li, "Spectral super-resolution network guided by intrinsic properties of hyperspectral imagery," *IEEE Trans. Image Process.*, vol. 30, pp. 7256–7265, 2021.

[42] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Pyramid fully convolutional network for hyperspectral and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 5, pp. 1549–1558, May 2019.

[43] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 517–532.

[44] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "a^2-Nets: Double attention networks," in *Proc. Int. Conf. Inf. Process. Syst.*, 2018.

[45] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, and F. Xu, "Compact generalized non-local network," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 6511–6520.

[46] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[47] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 371–381.

[48] Y. Qiu, R. Wang, D. Tao, and J. Cheng, "Embedded block residual network: A recursive restoration model for single-image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4180–4189.

[49] K. Liu and G. Mattyus, "Fast multiclass vehicle detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1938–1942, Sep. 2015.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Comput. Sci.*, 2014.

[51] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.

[52] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[53] Y. Guo et al., "Closed-loop matters: Dual regression networks for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5407–5416.

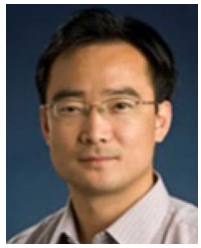
[54] S. Zhou, J. Zhang, W. Zuo, and C. C. Loy, "Cross-scale internal graph neural network for image super-resolution," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 3499–3509.
[55] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.
[56] Z. Pan, W. Ma, J. Guo, and B. Lei, "Super-resolution of single remote sensing image based on residual dense backprojection networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7918–7933, Oct. 2019.
[57] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
[58] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 63–79.
[59] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

Ziyu Liu received the B.S. degree in spatial information and digital technology, from the China University of Geosciences, Wuhan, China, in 2020 , where she is currently working toward the M.S. degree in computer science and technology.

Her research interests include high-resolution remote sensing image processing, remote sensing images super-resolution, and deep learning.

Ruyi Feng (Member, IEEE) received the B.S. degree in geographic information system from Hunan Normal University, Changsha, China, in 2011, and the M.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013 and 2016, respectively.

Since 2016, she has been with the School of Computer Science, China University of Geosciences, Wuhan, where she is currently an Associate Professor. Her research interests include sparse representation, deep learning, hyperspectral image analysis, high-resolution remote sensing understanding, and intelligent interpretation of remote sensing imagery.



Lizhe Wang (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1998 and 2001, respectively, and the D.E. degree (*magna cum laude*) in applied computing from the University of Karlsruhe, Karlsruhe, Germany, in 2007.

He is currently the Dean and the "ChuTian" Chair Professor with the School of Computer Science, China University of Geosciences, Wuhan, China. His research interests include remote sensing data processing, digital earth, and big data computing.

Dr. Wang is a Fellow of the Institution of Engineering and Technology and the British Computer Society. He was selected for Distinguished Young Scholars of the National Natural Science Foundation of China, the National Leading Talents of Science and Technology Innovation, and the 100-Talents Program of Chinese Academy of Sciences. He is an Associate Editor for *Remote Sensing*, *International Journal of Digital Earth*, *ACM Computing Surveys*, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, and IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING.

Tieyong Zeng received the B.S. degree in mathematics from Peking University, Beijing, China, the M.S. degree in applied mathematics from Ecole Polytechnique, Palaiseau, France, and the Ph.D. degree in mathematics from the University of Paris XIII, Paris, France, in 2000, 2004, and 2007, respectively.

He is currently an Associate Professor with the Department of Mathematics, The Chinese University of Hong Kong, Hong Kong. His research interests include image processing, optimization, artificial intelligence, scientific computing, computer vision, machine learning, and inverse problems.